BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problems Mailbox.

(3)

(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **10260991 A**

(43) Date of publication of application: **29.09.98**

(51) Int. Cl **G06F 17/30**

(21) Application number: **09270251**

(22) Date of filing: **02.10.97**

(30) Priority: **14.01.97 JP 09 5010**

(71) Applicant: **SEIKO EPSON CORP**

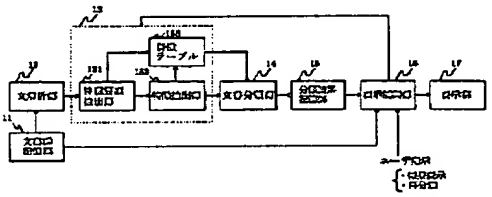
(72) Inventor: **MIWA SHINJI**

(54) **INFORMATION RETRIEVING METHOD AND INFORMATION RETRIEVING DEVICE**

(57) Abstract:

PROBLEM TO BE SOLVED: To accurately take out information that is needed for a user by watching the characteristic element of each displayed cluster to select a cluster and performing an operation, that interactively and also gradually narrows a retrieval object, of electing either resorting or result display.

SOLUTION: A characteristic extracting part 132 counts how many times which characteristic element appears based on characteristic elements that are extracted from a characteristic element extracting part 131 and generates a characteristic table 133. A document sorting part 14 refers to the table 133 and divides into plural clusters based on statistical information such as the appearance frequency of each characteristic element that exists in each document. A display controlling part 16 controls that the content of a sort result storing part 15 is shown as a sort result on a display part 17, also forms display data as a retrieval result content based on the content of the part 15 and the content of a document group storing part 11 when an instruction of retrieval result display is made by a user and shows the display data on the part 17.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平10－260991

(43)公開日　平成10年(1998)９月29日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ｄ
		15/40	３７０Ａ

審査請求　未請求　請求項の数６　ＯＬ　（全９頁）

(21)出願番号　特願平9－270251

(22)出願日　平成９年(1997)10月２日

(31)優先権主張番号　特願平9－5010

(32)優先日　平９(1997)１月14日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　000002369
セイコーエプソン株式会社
東京都新宿区西新宿２丁目４番１号

(72)発明者　三輪　真司
長野県諏訪市大和３丁目３番５号　セイコーエプソン株式会社内

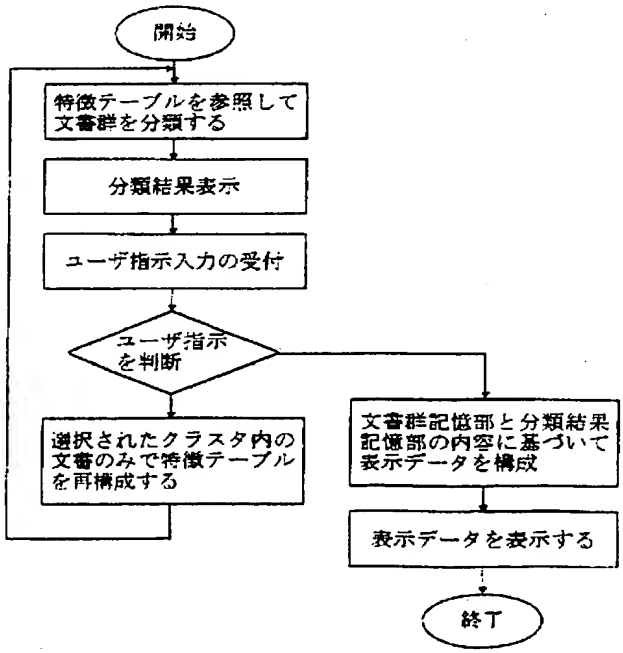
(74)代理人　弁理士　鈴木　喜三郎　（外２名）

(54)【発明の名称】　情報検索方法および情報検索装置

(57)【要約】

【課題】　適当なキーワードの指定が難しい場合でも的確な情報検索を可能とする。

【解決手段】　それぞれの文書から特徴要素を抽出し、その特徴要素とその特徴要素を含む文書との関係を表す特徴テーブルを作成して、その特徴テーブルを用いて文書群を複数のクラスタに分類して表示する（ステップｓ１，ｓ２）。このクラスタは、各クラスタごとにそのクラスタを代表する特徴要素とその特徴要素を一定以上含む文書数とをデータとして有する。そして、ユーザからのクラスタ選択指示を受けたとき、選択されたクラスタに属する文書内容の表示指示または再分類指示を受け付けて、内容表示指示の場合は、当該クラスタに属する文書内容の表示を行う（ステップｓ３～ｓ６）。一方、再分類指示の場合は、当該クラスタに属する文書のみで特徴テーブルを再構成して（ステップｓ３～ｓ５）、その再構成された特徴テーブルに基づいてクラスタに分類して表示する。



1

【特許請求の範囲】

【請求項１】　文書群に属するそれぞれの文書を解析し、それぞれの文書から特徴要素を抽出し、その特徴要素とその特徴要素を含む文書との関係を表す特徴テーブルを作成して、その特徴テーブルに基づいて文書群を複数のクラスタに分類して表示し、ユーザからのクラスタ選択指示を受けたとき、その選択されたクラスタに属する文書に関する内容の表示指示または再分類指示を受け付けて、内容表示指示の場合は、当該クラスタに属する文書に関する表示を行い、再分類指示の場合は、当該クラスタに属する文書のみで前記特徴テーブルを再構成してその再構成された特徴テーブルに基づいてクラスタに分類して表示することを特徴とする情報検索方法。

【請求項２】　前記特徴テーブルに基づいて文書群を複数のクラスタに分類する処理は、それぞれの文書内に存在するそれぞれの特徴要素の出現頻度などの統計的な情報に基づいて複数のクラスタに分類することを特徴とする請求項１記載の情報検索方法。

【請求項３】　前記分類された複数のクラスタは、少なくとも、それぞれのクラスタごとにそのクラスタを代表する特徴要素と、その特徴要素を一定以上含む文書数とをデータとして有することを特徴とする請求項２記載の情報検索方法。

【請求項４】　文書群を記憶する文書群記憶部と、
この文書群記憶部に記憶されているそれぞれの文書を解析する文解析部と、
この文解析部による解析結果からそれぞれの文書に対する特徴要素を抽出し、その特徴要素とその特徴要素を含む文書との関係を表す特徴テーブルを作成する特徴テーブル作成部と、
前記特徴テーブルの内容に基づいて文書群を複数のクラスタに分類する文書分類部と、
この文書分類部により分類された内容を記憶する分類結果記憶部と、
この分類結果記憶部の内容を読み出して複数のクラスタを表示させる制御を行うとともに、ユーザからのクラスタ選択指示を受けたとき、その選択されたクラスタに属する文書に関する内容の表示指示または再分類指示を受け付けて、内容表示指示の場合は、当該クラスタに属する文書を表示させる制御を行い、再分類指示の場合は、当該クラスタに属する文書のみで前記特徴テーブルを再構成させる制御を行う表示制御部と、
を有することを特徴とする情報検索装置。

【請求項５】　前記情報分類部が行う特徴テーブルに基づいてそれぞれの文書を複数のクラスタに分類する処理は、それぞれの文書内に存在するそれぞれの特徴要素の出現頻度などの統計的な情報に基づいて複数のクラスタに分類することを特徴とする請求項４記載の情報検索装置。

【請求項６】　前記分類された複数のクラスタは、少なくとも、それぞれのクラスタごとにそのクラスタを代表する特徴要素と、その特徴要素を一定以上含む文書数とをデータとして有することを特徴とする請求項５記載の情報検索装置。

2

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文書データを蓄積したデータベースやネットワークで公開されている文書群から、ユーザの要求に合致する文書を検索し、提示し得る情報検索方法および情報検索装置に関する。

【０００２】

【従来の技術】文書データを蓄積したデータベースなどからユーザの要求する文書を効率よく検索して取り出す方法としては、ユーザの入力したキーワードをもとに文書データを取り出すというような検索方法が一般的である。

【０００３】この検索方法は、検索しようとする文書に対してユーザ自身が何らかのキーワードを考えて、そのキーワードを入力することにより、システム側で、そのキーワードに合致する文書を取り出して出力するというものである。

【０００４】

【発明が解決しようとする課題】しかしながら、上述したような情報検索方法は、検索対象のデータベースの規模や格納されている文書の種類にかかわらず、入力されたキーワードのみによる検索であるため、状況によっては、検索されて出力される文書量が膨大な量となったり、逆に、検索結果が全く出ないということもある。これは、キーワードの設定仕方によっても大きく左右されるため、入力するキーワードを適切に選ぶことが必要となってくる。

【０００５】一般に、この種の検索方法においては、どのようなキーワードを入力したらよいかをユーザ自身で考える必要がある。たとえば、あるキーワードを入力したとき、所望とする文書が得られないような場合には、次に、違うキーワードを入力して検索してみるといった試行錯誤的な検索を行う必要がある。このように試行錯誤的にキーワードを入力して検索を行っても、ユーザが本当に必要としている情報が得られるとは限らない。

【０００６】また、情報をある程度絞り込む方法として、複数のキーワードを入力するという方法もあるが、ユーザ自身が何を取り出したらよいのかが明確にわかっていないような場合には、複数のキーワードを設定することは難しいし、また、複数のキーワードによって取り出される情報は、内容が絞り込まれ過ぎることもあり、かえって、所望とする情報を得にくくしてしまう場合もある。。

【０００７】ユーザはデータベースに保存されている内容全体を見ることは通常では不可能であるため、いわゆる情報検索という処理を行うわけである。しかし、現在

の情報検索は、データベース内にどのような情報が入っているかが殆どわからない状態で検索を行うために、何をキーワードとしたら最も適切な情報が取り出されるのかがわからないのが実情である。さらに、ユーザ自身、検索すべき情報に対して詳しい知識が無い状態で、どのような情報を得たらよいのか判断できないまま情報検索を行う場合もある。

【０００８】このような状況での情報検索を行う場合、従来のように、ユーザの入力したキーワードに基づいて検索を行う方式では、ユーザの所望とする情報を短時間で的確に得ることはできなかった。

【０００９】そこで本発明は、データベースの内容を段階的にアウトラインを示しながら表示し、ユーザはその表示を見て選択操作を行うことで、ユーザの必要とする情報を段階的に具体化していくことができるようにし、ユーザ自身がキーワードを考える必要がなく、また、ユーザ自身、検索すべき情報に対して詳しい知識が無い状態で、どのような情報を得たらよいのか判断できないまま情報検索を行う場合でも最終的にユーザの要求する情報を効率よく得ることができる情報検索方法および情報検索装置を提供することを目的としている。

【００１０】

【課題を解決するための手段】本発明の情報検索方法において、請求項１の発明では、文書群に属するそれぞれの文書を解析し、それぞれの文書から特徴要素を抽出し、その特徴要素とその特徴要素を含む文書との関係を表す特徴テーブルを作成して、その特徴テーブルに基づいて文書群を複数のクラスタに分類して表示し、ユーザからのクラスタ選択指示を受けたとき、その選択されたクラスタに属する文書に関する内容の表示指示または再分類指示を受け付けて、内容表示指示の場合は、当該クラスタに属する文書に関する表示を行い、再分類指示の場合は、当該クラスタに属する文書のみで前記特徴テーブルを再構成してその再構成された特徴テーブルに基づいてクラスタに分類して表示することを特徴としている。

【００１１】また、請求項２の発明は、請求項１の発明において、前記特徴テーブルに基づいて文書群を複数のクラスタに分類する処理は、それぞれの文書内に存在するそれぞれの特徴要素の出現頻度などの統計的な情報に基づいて複数のクラスタに分類するようにしている。

【００１２】さらに、請求項３の発明は、請求項２の発明において、前記分類された複数のクラスタは、少なくとも、それぞれのクラスタごとにそのクラスタを代表する特徴要素と、その特徴要素を一定以上含む文書数とをデータとして有している。

【００１３】また、本発明の情報検索装置において、請求項４の発明では、文書群を記憶する文書群記憶部と、この文書群記憶部に記憶されているそれぞれの文書を解析する文解析部と、この文解析部による解析結果からそれぞれの文書に対する特徴要素を抽出し、その特徴要素とその特徴要素を含む文書との関係を表す特徴テーブルを作成する特徴テーブル作成部と、前記特徴テーブルの内容に基づいて文書群を複数のクラスタに分類する文書分類部と、この文書分類部により分類された内容を記憶する分類結果記憶部と、この分類結果記憶部の内容を読み出して複数のクラスタを表示させる制御を行うとともに、ユーザからのクラスタ選択指示を受けたとき、その選択されたクラスタに属する文書に関する内容の表示指示または再分類指示を受け付けて、内容表示指示の場合は、当該クラスタに属する文書を表示させる制御を行い、再分類指示の場合は、当該クラスタに属する文書のみで前記特徴テーブルを再構成させる制御を行う表示制御部とを有することを特徴としている。

【００１４】また、請求項５の発明は、請求項４の発明において、前記文書分類部が行う特徴テーブルに基づいてそれぞれの文書を複数のクラスタに分類する処理は、それぞれの文書内に存在するそれぞれの特徴要素の出現頻度などの統計的な情報に基づいて複数のクラスタに分類するようにいている。

【００１５】さらに、請求項６の発明は、請求項５の発明において、前記分類された複数のクラスタは、少なくとも、それぞれのクラスタごとにそのクラスタを代表する特徴要素と、その特徴要素を一定以上含む文書数とをデータとして有している。本発明は、それぞれの文書内に存在するそれぞれの特徴要素の出現頻度などの統計的な情報に基づいて複数のクラスタに分類し、分類された複数のクラスタは、少なくとも、それぞれのクラスタごとにそのクラスタを代表する特徴要素と、その特徴要素を一定以上含む文書数とをデータとして有するようにし、これをユーザに表示するようにしている。

【００１６】これにより、ユーザは、表示された複数のクラスタのそれぞれの特徴要素をキーワードとして捉えることができ、それぞれのクラスタごとの特徴要素から、データベース内の概要を知ることができる。したがって、データベース内にどのような情報があるかが全くわからない状態で情報検索を行う場合でも、表示された複数のクラスタについて、ユーザが所望とするクラスタを選択し、かつ、選択されたクラスタ内の文書数が多すぎる場合には、再分類を要求するという処理を段階的かつ対話的に行うことによって、文書数の絞り込みが行え、絞り込まれた状態から最終的にユーザがクラスタを選択し、結果表示要求を行うことで、ユーザの所望とする情報を得ることができる。

【００１７】このように、本発明では、情報検索を行うに際して、表示されるクラスタごとの特徴要素を見てクラスタを選択する操作と、再分類か結果表示かを状況に応じて選択する操作を行えばよく、その操作過程でユーザの必要とする情報がどれであるかを段々と具体化して行くことができ、最終的にユーザの要求する情報を的確

に取り出すことができる。また、このような情報検索処理を行う過程において、入力すべきキーワードをユーザ自身が考える必要がなく、また、入力操作が選択操作ですむので、検索操作がきわめて容易なものとなる。

【００１８】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照して説明する。

【００１９】図１は本発明を実現するための装置構成を示す図であり、文書群記憶部１１、文解析部１２、特徴テーブル作成部１３、文書分類部１４、分類結果記憶部１５、表示制御部１６、表示部１７などから構成されている。

【００２０】文書群記憶部１１は、ある文書群に含まれる多数の文書をデータベースとして記憶するものである。

【００２１】たとえば、図２に示されるように、ある文書群として、「人工知能に関する論文群」があるとすると、その「人工知能に関する論文群」に属する論文として、たとえば、「エキスパートシステムに関する論文」、「自然言語処理に関する論文」、「ニューラルネットワークに関する論文」、・・・などがあり、さらに、たとえば、「エキスパートシステムに関する論文」の中には、「工場制御に関する論文」、「市場に関する論文」というように、ある１つの文書群には、多数の文書が存在している。

【００２２】文解析部１２は、文書群記憶部１１に記憶されているある文書群のそれぞれの文書データを基に、それぞれの文書の形態素解析を行い特徴要素としての単語を抽出する。

【００２３】特徴テーブル作成部１３は、特徴要素抽出部１３１、特徴抽出部１３２、特徴テーブル１３３から構成される。特徴要素抽出部１３１は、文解析部１２で形態素解析されて抽出された特徴要素（単語）を基に、それぞれの文書の中に存在する特徴的な単語を抽出する。特徴抽出部１３２は特徴要素抽出部１３１から抽出された特徴要素に基づいて、たとえば、それぞれの文書において、どの特徴要素が何回出現したかをカウントする。そして、特徴要素抽出部１３１から抽出された特徴要素と、特徴抽出部１３２でカウントされた数とにより、図３のような特徴テーブル１３３が作成される。

【００２４】図３に示される特徴テーブル１３３の例は、特徴要素としては、「エキスパート」、「システム」、「エキスパート・システム」、「自然」、「言語」、「自然・言語」が示されている。そして、「エキスパート」という特徴要素は、文書Ａには４回、文書Ｂには０回、文書Ｃには６回、文書Ｄには３回出現しており、また、「システム」という特徴要素は、文書Ａには４回、文書Ｂには０回、文書Ｃには８回、文書Ｄには５回出現しているというように、それぞれの特徴要素がそれぞれの文書にどのくらい出現しているかが示されている。

【００２５】この特徴テーブル１３３の内容によれば、文書Ａは、「エキスパート」や「システム」と言った特徴要素が多く出現し、文書Ｂは「自然」、「言語」、「自然・言語」といった特徴要素が多く出現し、文書Ｃは「エキスパート」、「システム」、「エキスパート・システム」といった特徴要素が多く出現し、また、文書Ｄは「エキスパート」、「システム」、「エキスパート・システム」、「自然」、「言語」、「自然・言語」などの特徴要素がどれも多く出現していることがわかる。

【００２６】文書分類部１４は、このような内容の特徴テーブル１３３を参照し、それぞれの文書内に存在するそれぞれの特徴要素の出現頻度などの統計的な情報に基づいて複数のクラスタに分類する。

【００２７】たとえば、１番目のクラスタとしては、そのクラスタを代表する特徴要素が「エキスパート・システム」であり、その「エキスパート・システム」という特徴要素を一定以上含む文書数は「２」であり、２番目のクラスタとしては、特徴要素が「自然・言語」であり、その「自然・言語」という特徴要素を一定以上含む文書数は「３」であるというように分類される。また、文書分類部１４は、このような特徴要素とその特徴要素を一定以上含む文書数の分類を行うとともに、それがどの文書であるかについての対応付けも行う。たとえば、文書数「２」に対応する文書は文書Ａと文書Ｃであるというような文書数と文書名の対応付けも行う。このような分類結果は分類結果記憶部１５に格納される。

【００２８】表示制御部１６は、分類結果記憶部１５の内容を分類結果として表示部１７に表示する制御を行うとともに、ユーザから検索結果表示の指示があったときは、分類結果記憶部１５の内容と前記文書群記憶部１１の内容に基づいて検索結果内容としての表示データを構成し、その表示データを表示部１７に表示する制御を行う。

【００２９】図４は表示部１７に表示された分類結果の一例を示すもので、この例では１つの画面上には、クラスタとしてたとえば分類１から分類１０まで、１０個のクラスタを表示する。なお、ここでは、１画面に１０個単位としたが、これは適当な数を設定できるものであり、また、クラスタ数が多い場合は、１０個ずつに分けてページ切替えで表示するようにすることも可能である。

【００３０】この図４の例では、分類１の特徴要素は「エキスパート・システム」であり、その文書数は「２」、分類２としては、特徴要素が「自然・言語」であり、その文書数は「３」であるというように表示されている。このように、各クラスタ毎にそのクラスタを代表する特徴要素とその特徴要素が一定以上存在する文書数が表示される。また、その表示部１７には「結果表示」と「再分類」といったユーザの指示を入力するため

のユーザ指示部２１、２２が表示される。

【００３１】ユーザはこのような表示内容を見て、ユーザ自身の要求している情報が、たとえば、分類１の内容（「エキスパート・システム」）に関係するものではないかと判断した場合は、その分類１の行部分Ｒ１をマウスなどでクリックしたのち、「結果表示」のユーザ指示部２１をクリックする。

【００３２】これにより、表示制御部１６は、選択されたクラスタ（分類１）に属する文書を文書群記憶部１１から読み出して、その文書内容を表示部１７に表示する。この例では、選択されたクラスタ（分類１）に属する文書数は「２」であり、その文書名が文書Ａと文書Ｃであることがわかるから、表示制御部１６は、ユーザからの結果表示要求を受けると、ユーザの選択したクラスタに属する文書（文書Ａと文書Ｃ）を文書群記憶部１１から読み出して、その内容を表示する。

【００３３】なお、この文書内容の表示の仕方としては、分類１に属するすべての文書（この場合文書Ａと文書Ｃ）の内容をそのまますべてを表示させるようにしてもよいが、たとえば、文書が論文である場合には概要を表す部分のみを表示するようにしてもよく、あるいは、文書名と文書サイズなどのみを表示するようにしてもよく、その表示の仕方については種々考えられる。

【００３４】一方、ユーザが図４に示すような表示内容を見て、ユーザ自身の要求している情報が、たとえば、分類１の内容（「エキスパート・システム」）に関するものではないかと判断したものの、「エキスパート・システム」という表示内容だけでは、ユーザ自身の要求する情報として具体化されていないと判断した場合、つまり、もう少し細分類化された内容が必要であると考えたときは、分類１の行部分Ｒ１をマウスなどでクリックしたのち、「再分類」のユーザ指示部２２をクリックする。

【００３５】このように、分類１の行部分Ｒ１がクリックされたのち、「再分類」のユーザ指示部２２がクリックされると、分類１に属する文書のみで再分類処理される。この図４に示す例では、分類１に属する文書数は「２」である。したがって、この２つの文書のみを用いて、それらの文書に存在する特徴要素に基づいて特徴テーブルを再構成する。つまり、この例で考えると、分類１に属する文書は文書Ａと文書Ｃであるから、これらの文書Ａと文書Ｃとで新たな特徴テーブル１３３が作成されることになる。そして、新たに作成された特徴テーブルを参照して、文書分類部１４が文書群の分類を行い、この文書Ａと文書Ｃのみについて分類されされた内容が表示部１７に表示される。

【００３６】このようにして、分類対象の文書が絞られた状態で、再分類された結果は、再分類前の特徴要素がある程度はそのまま出てくるが、新たな分類対象の文書間で見た場合、ある文書に特有の特徴要素が、分類結果

として出てくる場合もある。たとえば、分類対象の文書を文書Ａと文書Ｂとしたとき、文書Ｃでは「工場制御」という特徴要素の出現頻度が高いが、この「工場制御」という特徴要素は文書Ａでは殆ど出現しないという場合は、１つのクラスタとして、特徴要素が「工場制御」でその「工場要素」を含む文書数が「１」というような分類結果が表示されることになる。この表示例を図５に示す。図５では、分類３のクラスタにおいて、特徴要素が「工場制御」でその「工場制御」を一定以上含む文書数が「１」というように表示されている。

【００３７】そして、ユーザがその再分類された表示結果を見て、ユーザの要求する情報が「工場制御」に関する内容に近いと判断した場合には、図５における分類３の行部分Ｒ２をクリックし、かつ、「結果表示」のユーザ指示部２１をクリックすると、文書Ｃの内容が表示される。なお、この表示についても前記したように、対象となる文書の内容をそのまま表示させるようにしてもよいが、たとえば、文書が論文である場合には概要を表す部分のみを表示するようにしてもよく、あるいは、文書名と文書サイズなどのみを表示するようにしてもよく、その表示内容については種々考えられる。

【００３８】なお、以上の例は、説明を容易なものとするために、図４の段階で分類される文書数を「２」というようなきわめて少ない数で説明したが、実際には、図４の段階では、それぞれの分類における文書数は数百というような数となることもある。したがって、ユーザがたとえば、分類１のクラスタを選択し、かつ、再分類を指示すると、その分類１に属する数百の文書での再分類がなされ、その再分類された内容として、分類１における特徴要素とその文書数、分類２における特徴要素とその文書数、分類３における特徴要素とその文書数というように、それぞれの分類番号ごとにその特徴要素とその特徴要素を含む文書数が表示部１７に表示される。

【００３９】そして、ユーザが、その再分類されたそれぞれの分類番号に対する特徴要素を見て、ある分類番号のクラスタを選択し、かつ、再分類を要求すると、今度は、その選択した分類番号に属する文書だけでの再分類がなされる。たとえば、ユーザが、分類３のクラスタを選択し、かつ再分類を指示すると、その分類３に属する文書数での再分類がなされ、前記同様に、分類１における特徴要素とその文書数、分類２における特徴要素とその文書数、分類３における特徴要素とその文書数というように、それぞれの分類番号ごとにその特徴要素とその特徴要素を含む文書数が表示される。

【００４０】このような処理が繰り返し行われることにより、対象文書数が段階的に絞り込まれて行く。そして、文書数が絞り込まれた状態で、ユーザは表示された特徴要素を見て、最も適当と思われる特徴要素が表示されているクラスタ部分をクリックしたのち、「結果表示」のユーザ指示部２１をクリックする。

【００４１】これにより、たとえば、最終的な段階でユーザの選択した特徴要素を含む文書数が「２」であれば、その特徴要素を含む２つの文書の内容が表示されることになる。なお、この結果表示処理は、分類結果記憶部１５に記憶されている最新の分類結果内容と、文書群記憶部１１の内容を基に、対応する文書名が読み出され、検索結果となる表示データを構成して、その表示データを表示部１７に表示することにより行う。

【００４２】このように、本発明では、表示部１７に表示される分類結果（図４参照）における分類番号に対応する特徴要素がいわばキーワードとなるものである。

【００４３】したがって、ユーザは自分の要求する情報について、何をキーワードとしてよいかわからないような場合であっても、システム側で、ユーザの要求を具体化するための指標となる特徴要素を画面上に分類番号対応に表示し、さらに、その特徴要素を一定以上含む文書の数を表示するので、ユーザはデータベースの概要を知ることができ、また、ユーザ自身がキーワードを考える必要がなく、その表示内容を見て、選択するという対話的な操作が可能となる。

【００４４】そして、必要に応じて何段階かの再分類操作を経て文書数が絞り込まれたところで、最も適当と思われる特徴要素を選択してその結果表示を行うというような検索処理を行うことで、必要とする文書を的確に取り出すことができる。

【００４５】図６は以上説明したこの実施の形態の処理手順をフローチャートである。図６において、文書分類部１４が特徴テーブル１３３を参照して文書群の分類を行い（ステップｓ１）、その分類結果を表示する（ステップｓ２）。この分類結果の一例としては、たとえば、図４で示すような内容である。そして、ユーザがその表示を見て、「結果表示」か「再分類」かの入力を行うと、そのユーザ指示入力を受け付け（ステップｓ３）、ユーザの指示が結果表示か再分類かを判断する（ステップｓ４）。ユーザ指示が結果表示である場合には、文書群記憶部１１と分類結果記憶部１５のそれぞれの内容から表示データを構成して（ステップｓ５）、その表示データを表示部１７に表示する（ステップｓ６）。

【００４６】一方、ステップｓ４において、ユーザの指示が再分類である場合には、選択されたクラスタ内の文書のみで特徴テーブル１３３を再構成する（ステップｓ７）。そして、ステップｓ１に処理が戻り、再構成された特徴テーブルを用いて文書群を分類し、以下前記ステップｓ２～ｓ７の処理を行う。なお、指示されたクラスタ内の文書のみで特徴テーブル１３３を再構成する処理は、指示されたクラスタの文書が前述したように、たとえば、文書Ａと文書Ｃであるとすれば、この文書Ａと文書Ｃのみを用いて、それぞれの文書から抽出された特徴要素に基づいて特徴テーブル１３３を再構成する処理である。

【００４７】以上説明したように、この実施の形態によれば、特徴要素としての単語の出現頻度などの統計的情報によって、文書群を自動的に、所定の数のクラスタに分類するとともに、それぞれのクラスタを代表する特徴要素とその特徴要素を一定以上含む文書の数を抽出し、それらをユーザに見せることによって、ユーザは、データベースの概略を知ることができる。そして、このようにして分類されたクラスタについて、ユーザはそれぞれのクラスタごとの特徴要素をキーワードとして捉えて、最も適当と思われる特徴要素の存在するクラスタを選択する。

【００４８】このとき、選択したクラスタ内の文書数が多すぎる場合には、選択されたクラスタ内の文書のみで再分類して表示することが可能で、この再分類された表示を見て、その中で、最も適当と思われる特徴要素の存在するクラスタを選択するという操作を段階的かつ対話的に行う。これにより、文書数の絞り込みが行え、絞り込まれた状態から最終的にユーザがクラスタを選択し、結果表示要求を行うことで、ユーザの所望とする情報を得ることができる。

【００４９】なお、以上説明した実施の形態は、本発明の好適な実施の形態の一例であるが、本発明はこれに限定されるものではなく、本発明の用紙を逸脱しない範囲で種々変形実施可能となるものである。また、本発明の処理を行う処理プログラムは、フロッピィディスク、光ディスク、ハードディスクなどの記憶媒体に記憶させておくことができ、本発明は、それらの記憶媒体をも含むものであり、また、ネットワークからデータを得る形式でもよい。

【００５０】

【発明の効果】本発明によれば、それぞれの文書から特徴要素を抽出し、その特徴要素とその特徴要素を含む文書との関係を表す特徴テーブルを作成して、その特徴テーブルに基づいて文書群を複数のクラスタに分類して表示し、ユーザからのクラスタ選択指示を受けたとき、その選択されたクラスタに属する文書に関する内容の表示指示または再分類指示を受け付けて、内容表示指示の場合は、当該クラスタに属する文書に関する表示を行い、再分類指示の場合は、当該クラスタに属する文書のみで前記特徴テーブルを再構成してその再構成された特徴テーブルに基づいてクラスタに分類して表示するようにしている。これにより、ユーザは、表示された複数のクラスタのそれぞれの特徴要素をキーワードとして捉えることができ、それぞれのクラスタごとの特徴要素から、データベース内の概要を知ることができる。したがって、データベース内にどのような情報があるかが全くわからない状態で情報検索を行う場合でも、クラスタごとの特徴要素からデータベース内の概要を知ることができる。そして、分類された複数のクラスタは、少なくとも、それぞれのクラスタごとにそのクラスタを代表する特徴要

11

素と、その特徴要素を一定以上含む文書数とをデータとして有するので、表示された複数のクラスタについて、ユーザが所望とするクラスタを選択したとき、選択したクラスタ内の文書数が多すぎる場合には、再分類が可能で、この処理を段階的かつ対話的に行うことによって、文書数の絞り込みが行え、絞り込まれた状態からユーザが所望とする情報を得ることができる。

【００５１】このように、本発明では、情報検索を行うに際して、表示されるクラスタごとの特徴要素を見てクラスタを選択する操作と、再分類か結果表示かを選択するという対話的でかつ段階的に検索対象を絞って行く操作を行えばよく、その操作過程でユーザの欲する情報がどれであるかを段々と具体化して行くことができ、最終的にユーザの必要とする情報を的確に取り出すことができる。また、このような情報検索処理を行う過程において、入力すべきキーワードをユーザ自身が考える必要がなく、また、入力操作がマウスなどによる選択操作ですむので、検索操作がきわめて容易なものとなる。

【図面の簡単な説明】

【図１】本発明の実施の形態における情報検索装置の構成を示すブロック図。

【図２】本発明の実施の形態に用いられる文書群の例を示す図。

12

【図３】本発明の実施の形態における特徴テーブルの一例を示す図。

【図４】本発明の実施の形態における複数のクラスタ表示例を示す図。

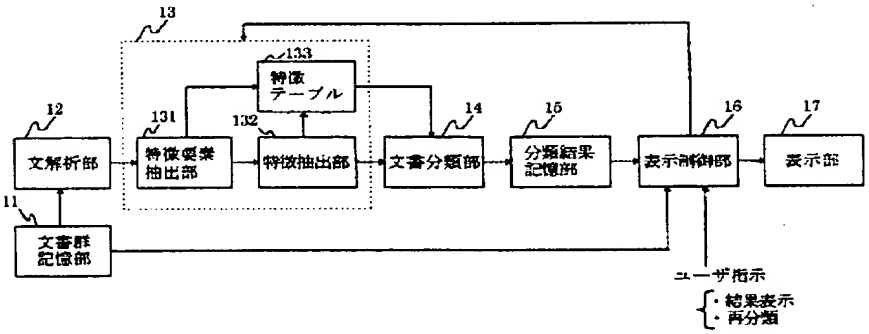
【図５】本発明の実施の形態における再分類指示により再構成された特徴テーブルに基づいてクラスタに分類された表示例を示す図。

【図６】本発明の実施の形態における検索処理手順を説明するフローチャート。

【符号の説明】

１１　文書群記憶部
１２　文解析部
１３　特徴テーブル作成部
１４　文書分類部
１５　分類結果記憶部
１６　表示制御部
１７　表示部
２１　「結果表示」のユーザ指示部
２２　「再分類」のユーザ指示部
１３１　特徴要素抽出部
１３２　特徴抽出部
１３３　特徴テーブル

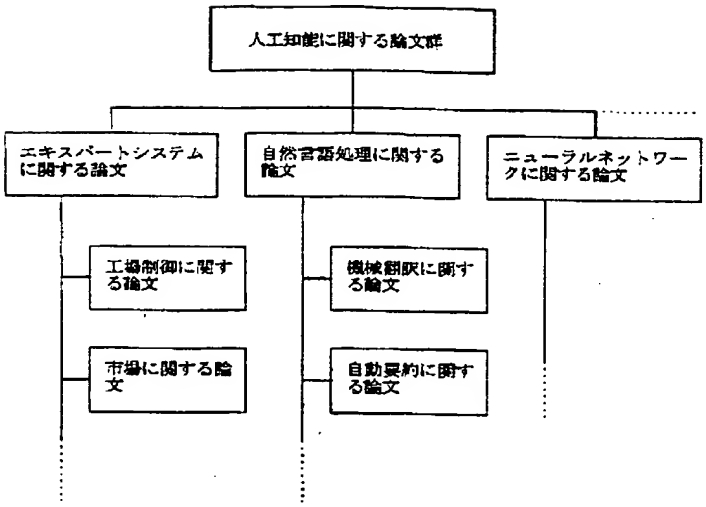
【図１】



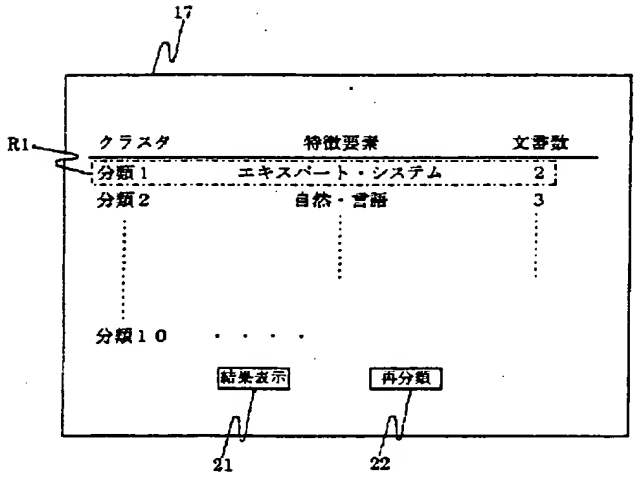
【図３】

特徴要素	文書A	文書B	文書C	文書D	…
エキスパート	4	0	6	3	…
システム	4	0	8	5	…
エキスパート・システム	2	0	4	2	…
自然	0	5	3	3	…
言語	1	6	0	3	…
自然・言語	0	5	0	3	…

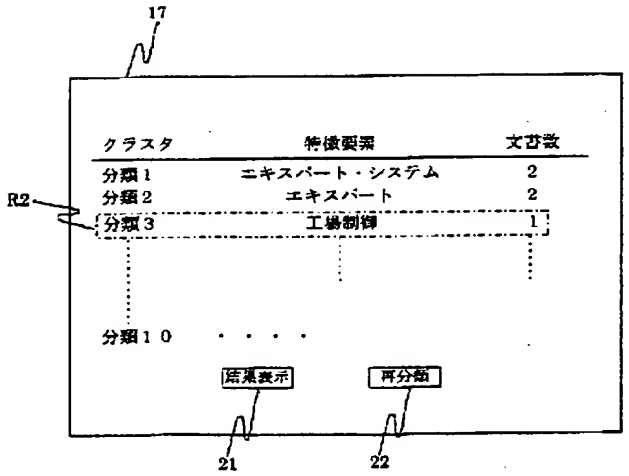
【図２】


人工知能に関する論文群
エキスパートシステムに関する論文
自然言語処理に関する論文
ニューラルネットワークに関する論文
工場制御に関する論文
市場に関する論文
機械翻訳に関する論文
自動要約に関する論文

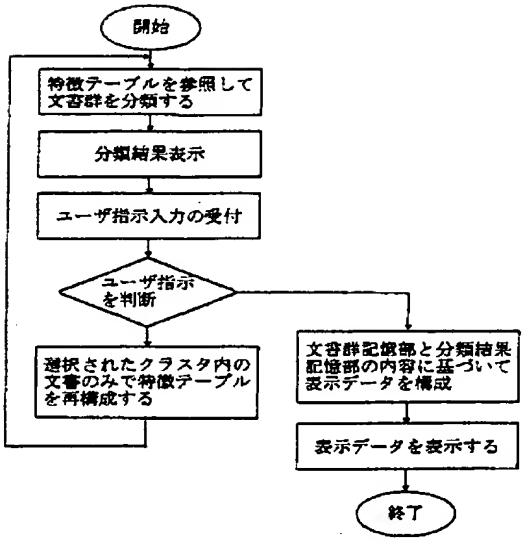
【図４】


17
R1
クラスタ
特徴要素
文書数
分類１
エキスパート・システム
2
分類２
自然・言語
3
分類１０
結果表示
再分類
21
22

【図５】


17
R2
クラスタ
特徴要素
文書数
分類１
エキスパート・システム
2
分類２
エキスパート
2
分類３
工場制御
1
分類１０
結果表示
再分類
21
22

【図６】


開始
特徴テーブルを参照して文書群を分類する
分類結果表示
ユーザ指示入力の受付
ユーザ指示を判断
選択されたクラスタ内の文書のみで特徴テーブルを再構成する
文書群記憶部と分類結果記憶部の内容に基づいて表示データを構成
表示データを表示する
終了